DOCUMENT RESUME

ED 138 590 SP 011 029

AUTHOR	Candor, Catherine
TITLE	Writing Classroom Tests.
INSTITUTION	Seattle Public Schools, Wash.; Washington Office of the State Superintendent of Public Instruction, Olympia.; Western Washington State Coll., Bellingham.
NOTE	26p.; Developed in conjunction with the TTT Writing Project Grant administered by the three agencies listed above
AVAILABLE FROM	Teacher Corps Project, Emporia Kansas State College, Emporia, Kansas 66801 (at cost)
EDRS PRICE	MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS	Educational Objectives; Evaluation Methods; Item Analysis; *Learning Modules; Mastery Tests; *Preservice Education; Questioning Techniques; *Teacher Education; *Test Construction
IDENTIFIERS	Item Classification

ABSTRACT

This document is a detailed module for instructing teacher trainees in the test development process. Upon completion of the package, the student is expected to be able to (1) identify pre-planning steps in test construction, (2) apply the two-way grid method of item classification, (3) develop and evaluate test items, and (4) apply item analysis to test results. Performance criteria are identified for each subtask involved in achieving the four terminal objectives, and appropriate exercises for mastering each subtask are provided. (MB)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED138590

WRITING CLASSROOM TESTS

Prepared by

Catherine Candor

Seattle Public Schools

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

EMPORIA KANSAS STATE COLLEGE
TEACHER CORPS
RESOURCE CENTER
ITEM NO. 943
FOR INSPECTION ONLY

(This module was developed in conjunction with the TTT Writing Project Grant administered by Seattle Public Schools in cooperation with Western Washington State College and the Office of the State Superintendent of Public Instruction.)

SP011 029

WRITING CLASSROOM TESTS

INTRODUCTION

Since all classroom teachers write tests at sometime or another, it is extremely important that they have the knowledge and skills necessary to write good tests. In this package you will select an instructional topic and follow through the test development process.

PREREQUISITE:

Defining Instructional Objectives.

RELATED PACKAGE:

Developing Criterion-Referenced Tests.

SUPPLEMENTARY TEXTS:

(recommended as resources)

Classroom Questions. What Kind? by Norris M. Sanders

From the introduction: Questions have always been the major stock in trade for teachers. At all levels, students are plied with different kinds of inquiries... This book should be of great assistance to those who make and ask classroom questions. Even more important, classroom questions will help students develop more effective and diversified thinking which should prove valuable not only in their student days, but, indeed, throughout their lives.

Measurement & Evaluation in Education by James Bradfield and H. Stewart Moredock

Fundamentals of Measurement by N. M. Downie

TERMINAL OBJECTIVES

Upon completion of this package you will be able to:

Task #I Identify pre-planning steps in test construction.

Task #II Apply a method of item classification, the two-way grid.

Task #III Develop and evaluate test items.

Task #IV Apply the statistics of item analysis to test results.

Task I. Identify pre-planning steps in test construction.

Performance Criterion #1

After reading the text below, the i.m. will be able to list the two major steps in test construction.

What is needed to write a good test? The answer is simple enough, but the actual job of writing is not so easy. Two major steps are required to write good tests: (1) specify the instructional objectives, and (2) write test items consistent with them.

For most teachers the major problem lies with specifying the instructional objectives. For many years it has been asserted in courses on teaching methods that good teaching requires a detailed list of instructional objectives, stated in terms of the behavior or competencies to be developed by students as the result of instruction. It appears, however, that in practice very few programs have clearly defined instructional objectives. If you have prepared instructional objectives, congratulations! You're half way home, free.

Otherwise, as a first step in test preparation secure all the instructional materials bearing on the intended test; for example, books, notes, and content outlines. Next you should resolve not to write any test items without first developing at least a simple plan for the test as a whole. Review the course materials and make statements of instructional objectives not necessarily in great detail, but at least in a general way. Since the content of the test helps define to the student what he is expected to have learned, this factor must be given very careful consideration. If you are unclear about how to state instructional objectives, even in a general way, do the package on "Defining Instructional Objectives" before you proceed.

Performance Criterion #2

After reading the following paragraph, the i.m. will be able to: (1) prepare a list of instructional objectives for material that he is currently teaching and wishes to "test," and (2) using this list of instructional objectives, answer the questions listed in the text.

After the instructional objectives are specified but before items are devised, the teacher must face up to such questions as:

(1) What kinds of student behavior would be evidence of progress toward each objective?

(2) What type of test situation and what specific items would elicit from students this type of behavior?

(3) How should such behavior be rated or scored?

(4) Is the stated objective realistic? If students have not met the objective, do I know how to reorient or modify instruction so as to obtain better results?

Performance Criterion #3

After reading the print context below, the i.m. will be able to paraphrase from memory the four criteria for a test designed to measure relative achievement. Give an example of how a test designed for instructional purposes could violate these criteria and still be a good test.

A critical variable is the purpose for which a test is to be used. Some teacher-made tests are used to measure the relative status of students in some aspect of achievement (the scores serving as a basis for assigning marks or otherwise ranking students). Others are designed chiefly to serve certain instructional purposes, identifying concepts that require reteaching, student self-evaluation, etc.

If a teacher is developing a test to aid in ranking students with respect to achievement in some area, the test should:

1. be based on a representative sampling of the content studied. The percentage of items on each topic should correspond approximately with the proportional emphasis given that topic in the course.

2. be based on a representative sampling of the abilities or skills emphasized in the course. Memorization of facts may be the only skill involved in a test unless the teacher makes a special effort to include questions that require the student to use higher level cognitive skills such as comprehension, application and analysis.

3. contain a sufficient number of questions so that the test will have adequate reliability. Although no exact rule can be given regarding number of items, a test used as a basis for grading should probably have a reliability coefficient of .70 or better. A measurement text should be consulted for procedures and formulas to determine reliability.

4. include items covering a wide range of difficulty. The test items should range from easy to difficult, with a large number of items geared to the middle group and with several items difficult to challenge the best student.

If the teacher is using his test for group diagnosis, student self-evaluation, concept mastery or other instructional purposes, and will not use the scores as a basis for ranking students, he need not be so concerned with the criteria listed above. For example, he may wish to include a disproportionate number of questions on a specific concept because of its complexity or because he thinks that reteaching of certain aspects may be needed. This disproportionate emphasis would institute a violation of criterion 1 and would distort the test scores as a basis for grading. The teacher may wish to emphasize certain skills to the exclusion of others (a violation of criterion 2). He may develop a short test that would be inappropriate as a basis for judging the achievement of individual students on a topic because of low reliability, but which would serve to show what facts or concepts had been learned by the whole group (see criterion 3). In a test on minimum essentials, e.g. a criterion-referenced post test, the teacher would not be concerned with range of difficulty of items; in fact, he might wish to include a large number of items that he hoped almost everyone had learned in order to assure himself that all students had mastered these basic concepts and to identify those who had not (see criterion 4). In short, an instructional test that was not designed to rank students on the basis of their total scores might violate one or more of the criteria listed above and still be a valuable classroom test.

Task II. Apply a method of item classification, the two-way grid.

Performance Criterion #1

After reading the following text, including that of Performance Criteria #2, the i.m. will be able to prepare a sample two-way grid covering material that he is currently teaching. Use the list of instructional objectives prepared in Task I, Criterion #2.

(1) Specify the purpose of the test, i.e. relative achievement or instructional.

(2) List the categories to be used along with the content or objectives dimension and the behavior or process dimension.

(3) Show how your two-way grid relates to your instructional objectives. (Satisfactio[n] of instructor and/or peer group.)

Ideally a teacher who is designing a comprehensive examination should define the universe to be sampled in much the same way as the author of a standardized test. That is, he should decide the proportional emphasis that should be given to different content areas; and he should also decide the types of abilities or behaviors to be sampled by the test items, e.g. recognition, recall, comprehension. That is, his plan for the test should consider the relative emphasis to be given both to content areas and to processes or cognitive abilities.

A two-way grid is a useful tool that can help a test writer prepare items consistent with his instructional objectives. It classifies each test item in accordance with two basic dimensions of each item - content and behavior. The number of items to be included under each classification should be decided and the grid prepared before the test items are written. This grid will help teachers overcome two major pitfalls of test writing:

(1) Over-emphasis of knowledge of specific facts and

(2) Over-emphasis of certain content areas in which items are easily constructed.

SAMPLE I

TEST ON MATHEMATICAL LOGIC

Objectives or Content	Behaviors			
	Exact Definition of Terms	Manipulation Skills	Reasoning	Total
Mathematical logic & proof	3	0	5	8
Properties of real-number system	9	3	7	19
Complex Numbers	2	4	2	8
Total	14	7	14	35

SAMPLE II

TEST ON ORIGIN OF EARTH (6th Grade Science)

Objectives or Content	Behaviors			
	Statement of Theory	Arguments for and against	"What would happen if"	Total
Early Theories	2	5	3	10
Present Theories	3	5	2	10
Total	5	10	5	20

Performance Criterion #2

After reading the paragraph below, the i.m. will be able to list the five levels of Bloom's Taxonomy and state a short definition of each.

The Taxonomy of Educational Objectives, by Benjamin S. Bloom provides a system of item classification which can be used along the behavior or process dimension of the grid. The Taxonomy provides for classification of educational goals and test items in the cognitive areas under the following five major categories:

1. Knowledge: involves the recognition and recall of facts and specifics.

2. Comprehension: the learner interprets, translates, summarizes, or paraphrases given material.

3. Application: involves the use of material in a situation which is different from the situation in which it was originally learned.

4. Analysis: involves separating a complex whole into its parts, until the relationship among the elements is made clear.

5. Synthesis: involves combining elements to form a new original entity.

The Taxonomy provides a logical and detailed system for item classification. Unfortunately it has not been widely used in planning tests - probably because of its complexity. There is also some danger in using the Taxonomy that the test writer will become too preoccupied with the task of writing items to fit the classification system (as, perhaps, when writing an item measuring synthesis). This concern may distract his attention from the main task of writing items that relate in a direct, commonsense way to his instructional objectives.

An example of how the Taxonomy can be used along the behaviors dimension follows.

SAMPLE III

FINAL EXAM IN NATURAL SCIENCE

Objectives or Content	Behaviors				
	Knowledge	Comprehension (Translation, Interpretation, Extrapolation)	Application	Analysis	Total
I. The Number Concept	6	4			10
II. Fundamental Concepts of Arithmetic	4	4	2		10
III. Quantitative Descriptions	2	3	3	2	10
IV. The Gas Laws	3	2	2	3	10
V. "The Spring of the Air"	1	2	3	4	10
VI. The kinetic Theory of Matter	4		4	2	10
VII. The Theory of the Atom	1	2	4	3	10
VIII. Electricity and Combustion	1		6	3	10
IX. Static Electricity and Magnets	2	2	3	3	10
X. Electricity and Nature of Matter	1	1	3	5	10
Total	25	20	30	25	100

Performance Criterion #3

After reading the material below, the i.m. will be able to classify test items along the behavior or process dimension according to Bloom's Taxonomy. If you feel you are already familiar with the material, take the self-test.

Ground Rules For Coding Cognitive Test Items*

In order to limit numerous interpretations and facilitate the classification of test items, Ground Rules are established. These ground rules assist in establishing reliability among coders.

1. When an item falls in more than one category, select the lower category.
2. If the answer to the item is given directly in the reading, it is coded Knowledge.
3. In addition and subtraction, if it is probable that the information is learned on a flash card (i.e., single-digit numbers, or no carrying or borrowing), it is coded Knowledge.
4. In multiplication and division, if the information is available in a 12 x 12 multiplication table, it is coded Knowledge.
5. In addition and subtraction, if it is necessary to carry numbers or borrow numbers for the correct answer, it is coded Application.
6. In multiplication and division, if the information is not available in a 12 x 12 multiplication table, it is coded Application.
7. Spelling is coded Knowledge.
8. Use of grammatical rules is coded Application.
9. Synonyms and antonyms are coded Knowledge.
10. Alphabetizing is coded Knowledge.
11. Following directions with visual clues is coded Knowledge.
12. Following directions without visual clues is coded Comprehension.
13. "What will happen" questions are coded Comprehension.

Cognitive Example*

Love is as important for a young child as the need for food. The child quickly becomes attached to his mother or to anyone who regularly cares for him. If she is away for even a few hours, he may show great anxiety, fearing that he has lost her. If she stays away for several days, he may be almost frantic with fear and grief. Unless she returns, or a loving substitute takes her place, his health may be seriously affected. Young children deprived of contact with other persons frequently sicken and may even die.

*Coding and Writing Test Items, Arizona: Educational Innovators Press, 1970.

Knowledge Item: The paragraph indicates that a child _____.

a. will readily become attached to physical objects.

b. will become attached to anyone who regularly cares for him.

c. often becomes more attached to his Father than his Mother.

d. likes food better than love.

Comprehension Item: Children depend upon their parents_____.

a. for emotional suport as much as for physiological nourishment.

b. for food only.

c. for money primarily

d. all their lives.

Analysis Item: The paragraph indicates that in the interest of raising an emotionally healthy child, you should_____.

a. occasionally "surprise" the child by leaving him alone.

b. try to be a reliable source of emotional support.

c. regularly punish the child so he won't think of of misbehaving.

d. regularly bring in a substitute mother for variety.

Application Item: The author seems to feel that_ _____.

a. in the case of young children, physical care is the equivalent of love.

b. children are sort of fickle about whom they regard as parents

c. children are independent and self-sufficient.

d. too much love can be harmful.

SELF-TEST*

Code the following test items according to the following Cognitive Levels:

K	Knowledge	An	Analysis
C	Comprehension	S	Synthesis
Ap	Application		

*Coding and Writing Test Items, Arizona: Educational Innovators Press, 1970.

When the level of cognitive behavior cannot be determined by definition, apply the appropriate ground rule by number.

1. Find the answer for each example:

_____ a. $\begin{array}{r} 59 \\ -54 \\ \hline \end{array}$

_____ b. 98 + 5 =

_____ c. $3\overline{)156}$

2. The children at my school play many games at recess. These are some of the problems:

_____ a. There are four places on the school grounds for playing ball. What else must you know in order to tell me how many can play ball at one time?

(1) how many pupils are in the room

(2) how many played last year

(3) how many like to play ball

(4) how many can play at each place

(5) nothing else

_____ b. If you know how many want to play ball and the number of ball diamonds, how many will play at each diamond? Choose one of these ways to find out.

(1) number of children take away number of diamonds

(2) number of diamonds multiplied by number of children

(3) number of diamonds plus number of children

(4) number of children divided by number of diamonds

(5) not given

Read the two paragraphs below. Decide which of the numbered words or phrases below the paragraph is best for each blank.

3. The Japanese, inhabitants of a group of islands off the China coast and close to Korea and Russia, have a long a . In fact, Jimmu, the first emperor, is said to have lived over 2500 years ago, and all succeeding emperors are said to be his descendents. Russian civilization did not reach the Pacific coast until recently, hence the greatest b influences on Japanese culture over the centuries have come from c .

_____ a. (1) winter
(2) coastline
(3) history
(4) super-highway

_____ b. (1) foreign
(2) emperor's
(3) Russian
(4) climactic

_____ c. (1) Russia and India
(2) China and Korea
(3) Russia and China
(4) Korea and Kuyshu

4. The distribution of living things depends on the distribution of the foods on which they live. Of all possible foods to sustain life, the most widely distributed is carbon dioxide, which occurs in air and is, therefore, to be found where a is. Thus the most widely distributed kind of b is that which is capable of feeding on carbon dioxide. The only type able to do this is that which produces the substance called cholorophyll. This fact provides a partial explanation of why plants are the c of all types of life on the earth.

_____ a. (1) water
(2) air
(3) food
(4) life

_____ b. (1) air
(2) carbon dioxide
(3) life
(4) food

_____ c. (1) most widely distributed
(2) greenest
(3) tallest
(4) most highly developed

5. Using oil paints, create a picture that you would hang in your room.

6. Circle the misspelled word in each group:

a.		b.	
	purfume		visited
	modern		changed
	divided		prize
	cries		amoung

7. Read the sentence, then decide which dictionary definition best defines the word as it is being used in the sentence. Look at the answer spaces at the right or on your answer sheet (if you have one). Fill in the space which has the same number as the definition you have chosen. Two other questions are asked about each word. For each of these, decide which answer is best and mark the space which has the same letters as the answer you have chosen.

for tune (for chən) n. 1. Great deal of money or property; riches; wealth. 2. Luck; chance; what happens. 3. Good luck; success; prosperity. 4. Fate; destiny; what is going to happen.

_____ a. He made his fortune trading in oil stocks.

_____ b. The gypsy told our fortune from the tea leaves.

_____ c. Fortune is the base (or root) word for all the following except:

(1) fortunate
(2) fortunately
(3) fortress
(4) misfortune

_____ d. In a dictionary, you would find fortune between the guide words:

(1) forty - fountain
(2) fortnight - fought
(3) formula - fortieth
(4) four - frank

8. Circle the correct verb form in each sentence:

(a) The Arabian Nights (are) (is) a famous book.

(b) Jimmy has (eaten) (ate) all the candy.

(c) The sun has (rose) (risen); it is time to get up.

9. Work the example. Then look at the possible answers at the right and see if your answer is given.

$67\overline{)6566}$

a. 87
b. 87-57/67
c. 88
d 88-7/67
e. not given

10. Read each question. Decide which of the answers given below is best.

_____ a. If a small community is looking for a doctor to come to its town, the most important thing to consider is whether or not he

(1) is a good doctor
(2) is married
(3) is a member of the community's church
(4) has a good family background

_____ b. Backwardness and odd ways of living are usually associated with:

(1) city dwellers
(2) mountain people
(3) valley people
(4) plains people.

11. Read each question. Decide which of the answers given below is best.

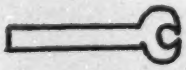
_____ a. Our class is planting seeds in two large boxes. The only difference is that box A will have earthworms and box B will not. We will probably find that the:

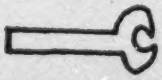
(1) plants in box A will be largest
(2) earthworms will eat the seeds in box A
(3) earthworms will eat the plants in box A
(4) plants in box B will be largest

_____ b. Which tool is best for loosening a very tight nut?

(1)

(2)

(3) 

(4) 

ANSWERS
(Cognitive Test Items)

Question	Answer	Ground Rule
1a.	K	3
1b.	Ap	5
1c.	Ap	6
2a.	An	Def
2b.	An	Def
3a.	C	Def
3b.	C	Def
3c.	An	Def
4a.	C	Def
4b.	C	Def
4c	C	Def
5.	S	Def
6.	K	7
7a.	C	Def
7b.	C	Def
7c.	C	Def
7d.	K	10
8.	Ap	8
9.	Ap	6
10a.	C	Def
10b.	C	Def
11a.	Ap	Def
11b.	Ap	Def

Task III. Develop and evaluate test items.

Performance Criterion #1

After reading the print context below, paraphrase a description of objective and essay tests, highlighting the differences between the two.

It is necessary here to distinguish between the objective and the essay test: the planning of each involves somewhat different problems.

The essay test, as opposed to the objective test, allows relative freedom of response to a given problem but at the same time requires the student to recall rather than recognize information, and to organize and express his ideas clearly and concisely. The test must to an extent be graded subjectively, and the scores assigned by various test graders may be inconsistent. Ordinarily an essay test consists of fewer questions than an objective test because of the time required to answer the questions.

There are many rules listed in various measurement texts regarding the design of essay questions. They will not be gone into in detail here. Rather three general suggestions for the use of essay tests are given below:

(1) Use essay questions to evaluate achievement of only those instructional objectives not as well or better tested by short-answer or objective items.

(2) Phrase the questions so as to require as precisely as possible the specific mental processes operating on specific instructional objectives.

(3) Essay questions must be stated in enough detail to give a clear understanding of what is expected in the answer, and points to be covered should be apparent

An objective test is so named because the system of scoring is objective rather than somewhat subjective as in the case of an essay test. Several possible responses are generally given for each test item with usually only one being the correct or best response. The test is scored by counting the number of times the person taking the test selected the correct responses from among those given. Presumably the result will be the same regardless of who scores the test. An objective test ordinarily consists of many more items than an essay test, and usually each item can be answered by making a single mark. The content of an objective test is subjective in the sense that it represents someone's judgment of what belongs in the test. It is the scoring that is objective.

These are some of the item types used in objective test writing:

Multiple choice item: This is the most common type of item. The correct answer for each test item is chosen from among several choices, usually four or five.

Matching item: This is a special kind of multiple choice item in which one set of options is used to answer (or match with) a number of different items.

True-false item: This is also a special kind of multiple choice item with only two options. There has been considerable objection to this kind of test item, principally because there is a fifty-fifty chance of guessing the correct answer. However, to meet this objection, the items can be adapted so that the student is required to indicate why a statement is false or to revise the statement so as to make it true.

Completion item: An item which requires the filling in or listing of one or two words is called a completion item and is usually classified as objective, although the scoring may not always be completely objective. Sometimes a subjective judgment has to be made as to whether the completion is adequate or accurate.

Performance Criterion #2

After reviewing the tables presented below, the i.m. will be able to state from memory at least three advantages and disadvantages for both essay and objective tests.

Tables I and II summarize the advantages and disadvantages of essay and objective tests. Those advantages or disadvantages that tend to characterize one approach or the other are listed first, followed by other factors that are greatly affected by the teacher's skill and care in test construction and scoring.

TABLE I

RELATIVE ADVANTAGES OF ESSAY AND OTHER SUPPLY-TYPE ITEMS***

Advantages	Disadvantages
1. Easily prepared a. Fewer questions to prepare b. Need not be mimeographed	1. May have relatively low reliability, owing to: a. Limited sampling of learnings b. Subjectivity of scoring
2. Largely eliminates guessing.	2. May have relatively low validity, owing to: a. Limited sampling of learnings b. Low reliability
3. Stimulates use of superior study methods in preparation (as compared with study methods used in preparing for objective tests).	3. Requires excessive time of students in writing
4.**Represents the most direct method of testing many outcomes. For some objectives, test exercises can closely approximate criterion behavior.	4.**Tends to be quickly and carelessly constructed, with the result that: a. Questions are ambiguously stated b. Questions are so general that a student can bluff or "talk around" the subject c. Questions are of unequal difficulty, with no provision for weighting them unevenly in the scoring process d. Selection of questions is not representative of major learnings.
5.**Provides more adequate basis for making inferences about student's level of competency; for example, his ability to define words, write clear explanations of procedures, or complete a geometric proof.	5.**Tends to be graded without an adequate scoring key, so that students' marks are affected by: a. The "halo effect" (good or bad) of a student's previous level of performance or success on a single question of the test b. Legibility of handwriting c. Errors of spelling and grammar d. Effectiveness of written expression.
6.**May give student opportunity to demonstrate his ability to: a. Choose most pertinent and important learnings b. Organize his knowledge c. Express opinions and attitudes d. Show initiative and originality	

**Variable factors which are markedly affected by the teacher's skill and care in test construction and scoring.

***From Measurement and Evaluation in Education, Psychology and Guidance, Georgia Sachs Adams. Holt, Rinehart and Winston, 1966.

TABLE II

RELATIVE ADVANTAGES AND DISADVANTAGES OF OBJECTIVE OR SELECTION-TYPE ITEMS***

1. Makes possible extensive sampling of learnings in a relatively short testing time.	1. Requires time and skill for adequate preparation. (Note: Good test questions, however, can be reused. Moreover, time spent by the teacher in test preparation may result in increased awareness of the goals of instruction, major concepts to be developed, facts related to such concepts, and the like.)
2. Tends to have high reliability as the result of: a. Objectivity of scoring b. Extensiveness of sampling	2. Is of limited value in some subjects.
3. Is scored objectively, with the result that: a. Scoring time is reduced b. The teacher is freed from suspicion of partiality c. Students' scores are not affected by such extraneous factors as ability to write rapidly and the like d. Students may do self-scoring e. Scoring may be delegated to clerical workers or readers f. Statistical analysis of student performance (on the test and on specific items) is facilitated g. At the upper grade levels, tests can be scored by machine and summaries made of student success on each test item.	3.**May stimulate superficial learning of many details because of: a. Emphasis placed on recognition of correct answer rather than on remembering b. Failure to require the student to organize significant facts and ideas and to reason about them.
4. Focuses student attention on specific facts or abilities being tested, permitting no evasion.	4.**May include ambiguous or misleading questions.
5.**May be valuable for such instructional purposes as: a. Pretesting b. Diagnostic testing c. Individualized self-testing d. Testing for application of principles to new situations	5.**May result in unduly high scores for intelligent, test-wise students who have not studied, because of: a. Transparent clues in grammar, word form, or phrasing b. Insufficient care in the devising of incorrect responses, allowing correct answers to be chosen more by the avoidance of obviously poor choices than by recognition of the correctness of the right response.

**Variable factors which are markedly affected by the teacher's skill and care in test construction and scoring.

***From Measurement and Evaluation in Education, Psychology and Guidance, Georgia Sachs Adams. Holt, Rinehart and Winston, 1966.

Performance Criterion #3

After reading the material below, the i.m. will be able to write and evaluate essay and objective test items.

Using the instructional objectives and the two-way grid developed in Task II Performance Criteria #1

(1) state whether each objective would be best evaluated by objective or essay test items.

(2) write appropriate test items. Use the two-way grid as your planning model (it may be modified if necessary).

(3) evaluate the test items and revise them if necessary (acceptance by peer group).

Note - Three very good sources of information about writing test items are: James Bradfield and H. Stewart Moredock, Measurement and Evaluation in Education; N. M. Downie, Fundamentals of Measurement; and Norris M. Sanders, Classroom Questions, What Kinds.

Much has been written about desirable characteristics of test items. Three of the most important are adequate content, clarity and the use of specific determiners.

Adequate Content: After completing the items, the test-maker has the responsibility of taking a careful second look at each one. He should ask himself: Is this item really worth including in the test? Does it cover pertinent material in an appropriate way? Does it measure important behaviors?

Clarity: Items should be stated clearly and in proper English. A prominent weakness in many test items is ambiguity--the meaning of a word, phrase, or sentence is not plain! Stereotyped expressions and textbook language should be avoided. Statements should not be "lifted" from the textbook.

Specific Determiners: A specific determiner is a clue to the answer inherent in the item but not relevant to the purpose of the item. Sometimes the way a test item is phrased will indicate the answer. For example:

(1) Avoid having the grammatical structure of the question or statement give a clue to the response. Use of the article "a" or "an," for example, will provide students with a clue.

(2) Do not give clues to the answer by varying the number or length of the blanks in a completion exercise.

(3) Avoid the use of words or phrases that are usually associated with a true or false statement. Such words as "all," "none," "always," "never" are usually associated with false statements; whereas statements containing "some," "generally," "may," "should" are usually true.

Test items should be written well in advance of the time that they are needed. The test writer may review his own item carefully and often will find flaws. He may ask other subject specialists to review his items and they may offer

suggestions for revision. In addition to revising ambiguous questions, the teacher should try to substitute improved options for any which are not plausible, revise the stem and/or options for any items incorrectly answered by a colleague, and eliminate or revise any items that appear to be extremely easy or difficult for the group. It is wise to write about 20% more items than will be included in the final version of the test. In addition a very useful review of the items will come from the students taking the test. They have a vested interest in getting the correct answer and will protest the loudest when the item is ambiguous or otherwise inadequate.

Since writing test questions requires considerable effort, it is recommended that over a period of time teachers build up a supply of items as a convenient source for tests. This means that as test items are written and tried out, they should be saved for future use.

Task IV. Apply the statistics of item analysis to test results.

Performance Criterion #1

After reading the text below, the i.m. will be able to compute difficulty level and discrimination index and interpret these statistics for given test results.

Responses to test items are analyzed statistically to provide information on how well or poorly the individual items and the test as a whole accomplish their purposes. Then what statistical characteristics should an item have? The answer depends on the purpose of the test. If the purpose is to measure relative achievement, then test items should yield certain results, as described below. On the other hand, if the purpose of the test is to measure a criterion behavior attainment of an instructional objective or other teacher standards, then the test writer will be interested in those statistical characteristics that reflect the extent to which the items have accomplished this purpose. Although the classroom teacher needs to be familiar with how test publishers use statistical data to select items, his primary concern should be with preparing good classroom tests and using test information to improve instruction.

In order to measure relative achievement, there must be enough difference among students' test scores so that true differences can be seen and compared; it is desirable that scores for a group of students be dispersed (spread out) as widely as possible. An adequate dispersion, or spread, of scores results when items are of (1) appropriate difficulty and (2) high discriminating power. These item characteristics can be determined through item analysis.

If the difficulty of the items contributes to the dispersion of scores, how difficult should the items be? In testing it has generally been found, other things being equal, that scores tend to be widely dispersed if many of the test items are at the middle level of difficulty; this means that many scores should fall halfway between a score which might be obtained by chance guessing and the highest possible score. For example, on an objective test of multiple choice items with five options, the correct answer might be guessed by chance one time in five, or 20% of the time. A perfect score would be 100% correct. In this case the middle difficulty level is halfway between 20 and 100% correct, which would be 60% correct. In order to disperse scores widely in a test with five choices per item, it is therefore desirable that many items be answered correctly by about 60% of the group.

In general, an item can be said to be discriminating in the positive sense if those who answer correctly tend to score higher on a criterion - what the test is supposed to measure - than those who answer incorrectly. The following example will help define a discriminating item: A particular item in a test was answered correctly by 75% of the students who did well on the test as a whole, scoring in the top fourth. On the same item, only 25% of the students who scored in the bottom fourth on the whole test answered correctly. Since those who were in the top fourth on total test score did well on the particular item in question and those in the bottom fourth did poorly, this item is said to be positively discriminating. If the situation were reversed, the item would be negatively discriminating. If the item does not differentiate at all between the groups, it is said to be nondiscriminating.

An index of discrimination--D. Various measures of item discrimination have been developed. A simple one is referred to as D. D is computed as follows:

$$D = \frac{\text{No. correct in upper 27\% -- No. correct in lower 27\%}}{\text{No. in 27\% of total group}}$$

To illustrate the application of the formula, assume that -

1. Ten students equal 27% of the total group. By the same token, there are 10 students in the upper 27% and 10 students in the lower 27%.

2. On a particular item, 9 of the 10 students constituting the upper 27% gave the correct response.

3. Five of the 10 students comprising the lower 27% gave the correct response.

$$D = \frac{9 - 5}{10} = \frac{4}{10} = .40$$

Generally speaking, items having a D of .30 or higher are considered satisfactorily discriminating and items as low as .20 are sometimes considered acceptable.

Computing item difficulty: The difficulty level is computed as follows:

$$\text{Difficulty level} = \frac{\text{(No. correct in upper 27\% + No. correct in lower 27\%)}}{\text{No. in upper 27\% + No. in lower 27\%}}$$

Example: Ten students equal 27% of the total group. On a particular item, 9 of the 10 students in the upper 27% and 5 of the 10 students in the lower 27% gave the correct response.

$$\text{Difficulty level} = \frac{9 + 5}{10 + 10} = \frac{14}{20} = .70$$

In computing both the index of discrimination and the difficulty level, the customary procedure is to select the highest and lowest 27% of the group, because this percentage has been found to work most effectively when results are analyzed for a fairly large group of students. However, teachers usually work with groups of 50 students or fewer; hence, it is best to compare results for the high-achieving and low-achieving halves of the group, because the results have greater reliability. The computational process remains the same.

The following is an example of how the difficulty level and the index of discrimination are applied to evaluate test items.

Use of Item Analysis in Planning Revision for Increased Discrimination

(High = Upper 27%; Low = Bottom 27%; 50 students in each group)

	A	B	C	D	E
Question I	11	*60 (41 high; 19 low)	15	2	12

Difficulty level = .60 D = .44

Item is discriminating. No revisions necessary.

Question II	20	5	32	*35 (18 high; 17 low)	8

Difficulty level = .35 D = .02

Item is poorly discriminating. Examine options C & D for possible ambiguities. Consider strengthening options B and E to make them more attractive.

Question III	*15 (15 high; 0 low)	44	21	11	9

Difficulty level = .15 D = .30

Item is discriminating but very difficult, possibly ambiguous. Check option B to be sure it could not be a correct answer; perhaps reword. Possible item is all right as is.

Question IV	*21 (10 high; 11 low)	21	23	23	12

Difficulty level = .21 D = .02

Item is poorly discriminating. Looks as though item is so difficult or ambiguous that everyone is guessing. Needs to be revised.

* Correct answer.

Self-Test

The item analysis data below is from an algebra-trigonometry test currently being developed by the Basic Skills-Mathematics Department of Seattle Public Schools to determine relative achievement of tenth and eleventh grade math students. Compute the difficulty level and the discrimination index. On the basis of these two statistics and the other data below, make suggestions for revision or review where appropriate.

Fourteen students in the upper half; fourteen students in the lower half; twenty-eight students in the total group.

	A	B	C	D	E
Question I	3	2	*17 (11 high; 6 low)	4	2
Question II	*3 (3 high; 0 low)	2	15	3	5
Question III	5	3	5	9	*6 (4 high; 2 low)
Question IV	*15 (10 high; 5 low)	4	3	3	2
Question V	4	*23 (13 high; 10 low)	1	0	0

* Indicates correct response.

Sample Answers to Self-Test

Question I — Difficulty level = .61 D = .35
Item is discriminating. Probably best to leave item as it is.

Question II — Difficulty level = .11 D = .21
Item is marginally discriminating and very difficult or ambiguous. Check option C to be sure it couldn't be a correct, perhaps reword. Needs to be revised.

Question III — Difficulty level = .21 D = .14
Item is poorly discriminating. Looks as though the item is so difficult or ambiguous that everyone is guessing. Needs to be revised.

Question IV — Difficulty level = .54 D = .35
Item is discriminating. Probably best to leave item as it is.

Question V — Difficulty level = .83 D = .21
Item is marginally discriminating. Consider strengthening options C, D and E to make them more attractive. Needs revision.

Performance Criterion #2

The discussion below relates to item analyses for tests designed for instructional purposes. Indicate whether the test you have developed (Task III, Performance Criteria #3) was designed to determine relative achievement or for instructional purposes. Then administer the test. For objective items compute the difficulty level and discrimination index. On the basis of these two statistics and other data gathered from the item analyses, make suggestions for revision or review.

The preceding discussion of item analysis presupposes that the major purpose of testing should be to measure relative achievement. In the case of classroom testing, however, a far more important purpose is to find out how a student has progressed in learning and what his next learning experience should be.

When test items are written to find out what Johnny knows, it is not important that their discrimination indices be high; in fact, in some cases the discrimination indices should be low because relative level of student achievement may be the same. The teacher is not trying to compare Johnny's performance with Joe's but to ascertain Johnny's grasp of classroom goals. When tests are used for this purpose, the item discrimination indices provide information about student performance and the effectiveness of instruction; they are not necessarily used to eliminate or revise items. A nondiscriminating item should, of course, alert the teacher to the possibility that the item is ambiguous and should be carefully examined. More importantly, however, statistical information can help in revising instructional plans. For example, a particular distractor in a multiple choice item may have attracted a substantial number of the better students. Why did this occur?

Was the item at fault, or was there ambiguity in the manner in which pertinent information was presented in the classroom? Careful examination of the item and discussion with students can provide some clues for course changes.

Information on item difficulty can also help determine how well students in general have achieved instructional goals, and can help identify the learning needs of individual students. The major objective of obtaining information on item difficulty is not to modify or eliminate items, but to provide this feedback.